IDENTIFICATION OF BACTERIA USING PHYLOGENETIC RELATIONSHIPS REVEALED BY MS/MS SEQUENCING OF TRYPTIC PEPTIDES DERIVED FROM CELLULAR PROTEINS

J. P. Dworzanski*
Geo-Centers, Inc., Aberdeen Proving Ground, MD 21010-0068

C. H. Wick and A. P. Snyder
U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010-5424

S. V. Deshpande
Science and Technology Corporation, Edgewood, MD 21040

R. Chen and L. Li
Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2

ABSTRACT

Genomes of all priority bacterial pathogens for biodefense purposes and more than one hundred other bacteria have been sequenced. In addition four hundred bacterial genome-sequencing projects are in progress. These achievements provide new possibilities for reliable identification of bacteria on a molecular level by retrieving their genomic information. We present a mass spectrometric approach to link these resources with genomic information encoded in bacterial proteins. Amino acid sequencing of tryptic peptides is performed from bacterial protein extracts. The results of searching tandem mass spectra of peptide ions against a comprehensive bacterial proteome database are analyzed using probability based scoring. This is followed by bacterial identification with an algorithm that uses phylogenetic relationships between bacterial species as a part of a hierarchical decision tree process.

1. INTRODUCTION

The detection and identification of pathogenic agents of biological origin including viruses, bacteria and toxins play a crucial role in a proper response to unintentional or terrorist caused outbreaks of infectious diseases and the use of biological warfare agents on the battlefield. Recently, genomes of more than one hundred bacteria, including those listed as priority bacterial pathogens for biodefense purposes have been fully sequenced (http://www.ncbi.nlm.gov/PMGifs/Genomes/micr/html). Moreover, almost five hundred other sequencing projects are in progress. This growing number of completely sequenced bacterial genomes, provides the sequence information of every potentially expressed protein encoded by these organisms (Blattner et al., 1997). The combination of this unprecedented resource with advances in mass spectrometry (MS) technologies capable of identifying proteins that are actually expressed (Aebersold and Goodlett, 2001) enables the design of new molecular diagnostic procedures to study the relatedness and identity of microorganisms. Our goal was to develop a method for bacteria classification and identification based on amino acid sequence information of their proteins by tandem mass spectrometry (MS/MS) analysis of peptide ions.

Classification and identification of microorganisms based on relationships encoded in DNA sequences is a commonly used approach (Pershing et al., 2003). However, the direct retrieval of sequencing information by full genome sequencing is time consuming and not practical for fast detection and identification purposes needed by our armed forces and homeland security requirements. Hence, there is an urgent need to develop fast and reliable methods to retrieve parts of genomic information that are thought to be representative of the whole genome.

Several groups have reported the application of MS/MS to obtain partial protein sequence information for the purpose of microorganism identification. (Chen et al., 2001; Harris and Reilly, 2002; Yao et al., 2002). More recently, Warscheid et al. (2003) demonstrated that the acid extraction of bacterial spore proteins followed by peptide micro-sequencing, using matrix assisted laser desorption/ionization (MALDI)-MS/MS data has the capability to identify species from the genus *Bacillus* in spore mixtures.

We have developed a bioinformatics procedure based on analysis of an electrospray ionization (ESI)-MS/MS data for the fast classification of analyzed bacteria, using phylogenetic relationships among them revealed by amino

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Identification Of Bacteria Using Phylogenetic Relationships Revealed By Ms/Ms Sequencing Of Tryptic Peptides Derived From Cellular Proteins	**g**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Geo-Centers, Inc., Aberdeen Proving Ground, MD 21010-0068; U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010-5424**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**8**	

acid sequences of their proteins. In short, proteins released during cell lysis are trypsin digested and generated peptides are analyzed using one-dimensional (1D) reversed-phase (RP)-liquid chromatography (LC) and ESI-MS/MS to obtain product ion mass spectra of peptide ions. These spectra are analyzed by a computer search algorithm to compare them with theoretical spectra generated *in silico* from peptide sequence strings in a protein database. We have constructed a bacterial proteome database dedicated to bacterial identification which contain amino acid sequences translated from all putative protein coding open reading frames (ORFs) of all sequenced bacterial genomes. To facilitate the process of bacterial identification based on product ion mass spectra searching results, we have also developed a statistical scoring algorithm to rank the validity of a sequence assignment by a searching engine (Dworzanski et al., 2004). To classify bacteria we examined sequence assignments to proteomes of bacterial taxons using phylogenetic relationships of database organisms. Our identification algorithm uses assignments of analyzed organism(s) to taxonomic groups based on an organized scheme that begins at the phylum (division) level and follows through classes, orders, families and genera down to the strain level. Finally, to identify a bacterium we applied hierarchical clustering algorithm to reveal relatedness of the analyzed expressed gene products to closest matches in the bacterial database. The applicability and reliability of this algorithm, called ABOid©, for identification of bacteria is demonstrated.

2. MATERIALS AND METHODS

2.1. Bacterial Samples.

Escherichia coli K-12 (ATCC 47076), *Bacillus subtilis* (ATCC 23857), *B. thuringiensis* (ATCC 33679), *B. cereus* (ATCC 14579), *Agrobacterium tumefaciens* (ATCC 33970) *and Lactococcus lactis* (ATCC 11454) were ordered from the American Type Culture Collection. Bacterial cells were incubated under ATCC recommended conditions, harvested, washed with distilled water, lyophilized, and stored at -25°C.

2.2 Cell Lysis, Protein Digestion, and Sample Clean up.

Proteins were extracted from bacterial cells after lysis with sonication (Branson probe sonicator; Branson Ultrasonics Corp., Danbury, CT) in 100 mM ammonium bicarbonate buffer (pH 8.5) for 2 minutes (1 pulse per second with 0.75 s pulse duration). The resulting suspension was centrifuged at 11,750 *g*. The supernatant was then filtered using Microcon-3 filters (Millipore, Mississauga, ON) with a 3000 Da molecular mass cut-off.

The cell extract was denatured with urea, reduced with DTT, alkylated by IAM, and digested by trypsin at a ratio of 1:50 (w/w). A Zip Tip C18 was used to desalt the resulting peptide mixture (Millipore, Mississauga, ON) before HPLC analysis.

2.3. HPLC and Acquisition of Mass and Tandem Mass Spectra.

1D HPLC-MS/MS was conducted on an LCQ DECA Surveyor LC-MS system (ThermoFinnigan, San Jose, CA). Chromatographic separation was performed on a Vydac C18 column (300 Å, 5 μm, 150 μm i.d. × 150 mm) with a flow rate of 1 μL/min. The mobile phase consisted of water and MeCN, and both contained 0.5% (v/v) acetic acid.

For two-dimensional (2D) HPLC-MS/MS, the peptide mixtures were first separated on a Vydac sulfonic acid cation-exchange column (900 Å, 8 μm, 300 μm i.d. × 150 mm) using a step-gradient with increasing NaCl concentration from 0 to 500 mM. Solvent delivery was performed on an Agilent (Palo Alto, CA) HP 1100 HPLC system. Two alternating Vydac C18 columns in an automated fashion separated the effluent from the first dimension Vydac column.

The LCQ DECA ion trap was operated using the Instrument Method files of Xcalibur and some of the key operational parameters of this instrument are listed below. The LCQ was set to acquire a full-scan mass spectrum between m/z 400 and 1400 followed by three data-dependent product ion mass spectral scans between m/z 400 and 2000 of the most intense precursor ions. The excitation energy for the precursor ions selected for collision-induced dissociation (CID) was set as 35% (using the operational parameter "% relative collision energy") with a 30 ms activation time. To avoid the collection of same ion spectra during a specified time period a method of acquisition termed dynamic exclusion was used with the following parameters specified by a ThermoFinnigan software: a repeat duration of 0.5 min, repeat count of 2, and a 3 min exclusion duration window.

2.4. Proteome Database.

A proteome database was constructed using the annotated bacterial proteomes or the raw genome sequences. Genome sequences were downloaded via the Internet from the National Center for Biotechnology Information (NCBI) site in a FASTA format and were automatically processed. A computational Gene Locator and Interpolated Markov Modeler (GLIMMER 2.02), made available by The Institute for Genomic Research (TIGR, Rockville, MD) was used to recognize protein-coding sequences (Saltzberg et al., 1998). In-house written software was applied for automatic translation of these codons into amino acid sequences of all putative proteins and for assembling a proteome database from the

available bacterial genomes. The proteome database was searched directly or was additionally processed to create a peptide sequence database which contains all potential tryptic peptides derived from all proteomes using a TurboSEQUEST utility (ThermoFinnigan). These translated sequences of all bacterial peptides were stored and indexed in a file that can be read and searched by product ion mass spectra data mining software such as SEQUEST. The 496,756 protein-encoding putative genes recognized by gene finding algorithms were used for the database construction. In some special cases of data processing unique matches between predicted peptide sequences and a given bacterium in our database were additionally analyzed to identify known gene products using BLASTP.

2.5. Data Processing.

The results of searches of product ion mass spectra against the protein database using the SEQUEST algorithm (Eng et al., 1994) were processed using an in-house developed new software application called ABOid©. This software was developed using Dynamic Programming Techniques and represents a Microsoft.NET front-end application integrated with a back-end developed in Oracle 9i to store the bacterial information. It also communicates with the computational dynamic linked libraries (DLL) engine written in Java to allow for easy portability to any hardware platform and making it platform independent. This software implements analyses based on a true probabilistic system for the evaluation of sequence assignments and provides functionalities that allow for automated classification and identification of bacteria as described in more details below.

The SEQUEST database search engine produces a list of peptide sequences with several matching parameters, which provide a relative indication of the validity of a sequence assignment to an MS/MS spectrum. Five parameters, namely Xcorr, ΔCn, Sp, RSp, and ΔMpep (see detailed descriptions of these terms in the Results and Discussion), were used in this work to arrive at a unified matching score and to compute a probability that the amino acid sequence assignment is correct. To train the system and to determine the unified matching score, discriminant function analysis was performed on the five parameters using a training dataset consisting of 3,019 peptide ion MS/MS spectra generated from an *E. coli* K-12 protein extract. In this case, product ion mass spectra were obtained by reserved-phase HPLC-MS/MS analyses of 17 ion-exchange-LC separated fractions from an *E. coli* K-12 digest. Each spectrum was searched using SEQUEST against the proteome database. If the top peptide candidate listed in the searching result was from a putative protein of *E. coli* K-12, this result was considered as a correct assignment while assignments to peptide sequences matched to proteomes from other bacteria as incorrect assignments. Discriminant analysis and modeling of discriminant function score distributions among correct and incorrect peptide assignments were then performed using Statistica software (release 6, StatSoft, Inc., Tulsa, OK) with the five matching parameters as variables.

3. RESULTS AND DISCUSSION

The 1D LC-MS/MS analyses (45-60 min) of tryptic digests of proteins derived from pure cultures and bacterial mixtures of model microorganisms were initially processed using SEQUEST. Fig. 1 shows the schematic representation of this proteomic approach for identification of bacteria based on MS/MS analysis of a whole cell protein digests, database search and statistical analysis of the matching scores. To obtain product ion mass spectra of tryptic peptides, cellular proteins were extracted and digested with trypsin, and HPLC-MS/MS analysis was performed. In this work, product ion mass spectra are searched against a database composed of genome-translated proteomes of bacteria. The searching results are analyzed using in-house developed software for statistical scoring of peptide assignments and bacterial identities. The key components of the method presented in Fig. 1 are described below including the validation of the data analysis procedure using testing samples containing one or a mixture of bacteria.

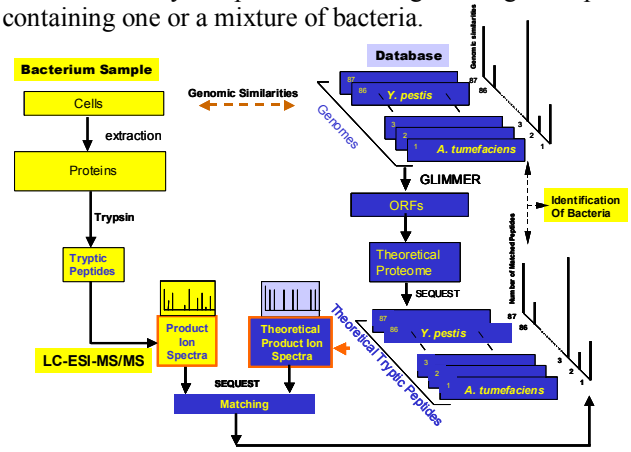

Fig. 1. Schematic representation of the experimental setup and data processing used for the identification of bacteria.

3.1 Database-Match Scoring.

Hundreds of product ion mass spectra of peptides generated from a bacterial cell protein extract digest at the rate of ca. 1 spectrum per second were searched against the proteome database by using a SEQUEST algorithm. An example of the product ion spectrum recorded during analysis of *B. cereus* is shown in Fig. 2 and is accompanied by an amino acid sequence of a peptide ion revealed by analysis of its fragmentation pattern.

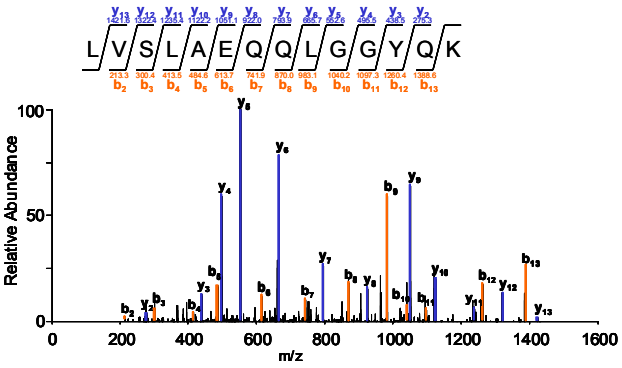

Fig. 2. Representative product ion mass spectrum from the doubly charged ion at m/z 768 derived from a tryptic peptide of *B. cereus* proteins. Labels show the most informative fragments that are interpreted above. SEQUEST generated matching parameters for the assignment: Xcorr=5.12; ΔCn =0.47; Sp= 2292.3; RSp= 1.0; ΔM_{pep} =0.72.

In this case the identified peptide was matched to the small acid soluble protein that play a protective role for DNA inside spores. Although all sporulating bacteria produce such proteins the revealed amino acid sequence is unique only for *B. cereus* strain 14789. Thus suggesting the possible identity of an analyzed organism.

Although SEQUEST may provide assignments of peptides to bacterial proteins or proteomes automatically, the validity of each match has to be evaluated to distinguish between real hits (i.e., correct assignments) and random matches (i.e., incorrect assignments). Moreover, for high throughput analysis of peptide assignments to bacterial proteomes the discrimination between correctly and incorrectly assigned peptides has to be performed automatically. To this end, our newly developed software ABOid© models SEQUEST computed scores using discriminant function (DF) analysis. The SEQUEST-generated scores represent variables associated with the quality of fit between an MS/MS mass spectrum and theoretical spectra generated from all possible tryptic peptides derived from protein sequences of all bacteria in the database. Multivariate DF analysis (Keller et al., 2002) maximizes the ratio of between-class variance to within-class variance by calculating appropriate weights associated with each variable and transforms SEQUEST scores into DF scores. The model was used to support the selection of peptides that would be subsequently used for analysis of sequence-based similarities between the analyzed test sample and microorganisms in the database.

The variables used for the multivariate DF analysis include a raw cross-correlation score of the top candidate peptide (Xcorr), the difference or delta in cross-correlation score (normalized to the highest Xcorr value) between the top-ranked peptide sequences (ΔCn), and the preliminary score of the top candidate peptide (Sp) used to rank the top 500 matching sequences. Additional variables contributing to discrimination comprised the natural logarithm of the rank of the preliminary raw score Sp (RSp) among the candidate peptides and the absolute value of the mass difference between a peptide characterized by a molecular ion with a postulated charge state and the theoretical mass of the assigned peptide (ΔM_{pep}).

To carry out the DF analysis, a large number of MS/MS peptide spectra from a known bacterium are required as a training dataset. To this end, a 2D HPLC-MS/MS analysis of an *E. coli* K-12 tryptic digest was performed and a total of 3019 MS/MS spectra were generated. These spectra were searched against the database by SEQUEST and the resulting matching data were divided into correct and incorrect assignments. The model was generated by fitting distributions of DF scores for the incorrect and correct peptide assignments using log-normal and Gaussian distribution functions, respectively. A user selected decision criterion for accepting peptides for further analysis is associated with the tradeoff between the sensitivity (i.e., the fraction of peptides belonging to the correct assignments to be used for bacterial identification) and the error rate, i.e., the fraction of peptides belonging to the incorrect assignments to be used for bacterial identification. To gauge the probability of correct assignment of a peptide from an unknown sample MS/MS spectrum to a microorganism proteome, statistical evaluation of the spectral matching can be carried out. In this work, DF score distributions experimentally observed and approximated using modeling functions were used to calculate the observed and expected probabilities that a peptide is correctly identified.

The results of the analysis are presented in Fig. 3 in the form of a probability curve, which indicates how DF scores can be translated into the probability of correct peptide assignments. In addition to the probability curve, Fig. 3 also shows a plot of the fraction of peptides with true positive assignments (sensitivity) as a function of the DF score and a plot of the fraction of peptides with false positive assignment (error) as a function of the DF score. It can be seen from Fig. 3 that, as the probability of correct assignment increases, the fraction of peptides with true positive assignments and the fraction of peptides with false positive assignments both decrease. However, they decrease at different rates. It should be noted that since only peptides with true positive assignments to the proteome of a microorganism contribute to the identification of this microorganism, a highly sensitive method should detect as many peptides with true positive assignments as possible. However, there are both biological and methodological reasons that can cause poor

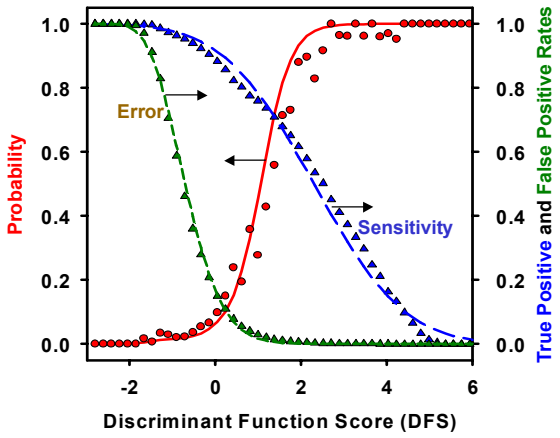

Fig. 3. Probability of correct peptide assignments and the fractions of correct (sensitivity) and incorrect peptide assignments (error) at different DF threshold scores.

matches between recorded MS/MS spectra and the respective theoretical spectra of peptides and give lower DF scores. The biological reasons include unsuspected post-translational modifications or single nucleotide polymorphism as well as other mutations specific for a given strain and expressed in the investigated proteome. The methodological errors include those associated with the preparation of the database (e.g., DNA sequencing, gene finding) manifested by the lack of a given sequence in the database and those related to the sample processing procedure.

3.2 Statistical Scoring for Bacterial Identification

As depicted in Fig. 1 the MS/MS spectra obtained from a 1D HPLC MS/MS experiment are searched against the proteome database by using SEQUEST, and the searching results are analyzed by ABOid© to calculate probabilities of correct amino acid sequence assignments. Only assignments passing the user determined probability level of correctness are further analyzed. Sequences of such peptides are used to match them to proteomes of a given taxon (phylum, class, order, genus or species). For analysis of *E. coli* derived peptides the distribution of the number of accepted peptide sequences to proteomes of bacterial species assembled in the small database of 87 organisms is shown in Fig. 4A. This figure displays a histogram with each bar representing the number of accepted peptides assigned to an individual database bacterium. In this case, the probability threshold used to accept peptides corresponds to 80%, hence, the maximum number of incorrect matches do not exceed 20%. In this case the number of unique (U) accepted peptides equals 171 with 80% of them correctly assigned to the K-12 strain. However, the real number of peptides matched to *E. coli* K-12 is higher than indicated above. This discrepancy originates from the presence of paralogous proteins, i.e., products of related genes formed by a gene duplication event. Moreover, note that a given peptide may match with several proteomes of different bacteria due to the presence of ortologous proteins sharing segments with the same amino acid sequences. It is assumed that although such proteins are expressed in separate species they are encoded by genes that are derived from the common ancestor or are products of the horizontal gene transfer between different strains.

As a consequence, these degenerate peptides are not uncommon and may be used to reflect genomic similarities between database organisms. For example, in Fig. 4A, the total number of all assigned peptides (T) is more than 6 times higher than the number of unique peptides accepted (U) because in a case of closely related organisms, like different *E. coli* strains, the core of expressed proteins is almost identical. However, each of these peptides, including those derived from ortologous proteins, differ in their power for bacteria discrimination. For example, sequences of peptides derived from a periplasmic binding protein that is involved in oligopeptide transport were found only in proteomes of *E. coli* and *S. flexneri* strains, while one of them (SGEIDMTNNSMPIELFQK) was found exclusively in the proteome of the K-12 strain. On the other hand, peptides derived from house keeping enzymes like isocitrate lyase do not provide discrimination between *E. coli* strains.

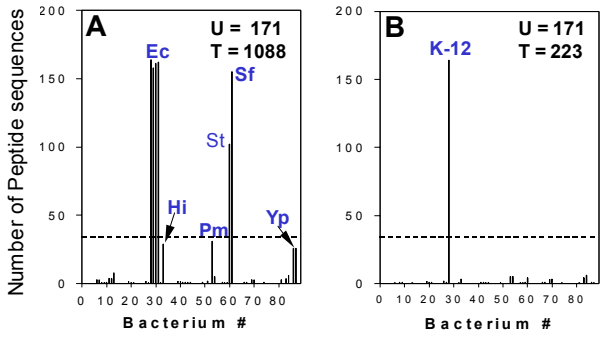

Fig. 4. Number of accepted peptides from the analysis of an *E. coli* K-12 protein digest matched to proteomes of bacteria in the database [Bacterium number, where consecutive numbers refer to the alphabetic order of bacterial genomes in the database]. (A) Peptide sequence assignments passing the filter at the 80% probability level; (B) the corresponding identification plot obtained by filtering out the degenerate sequences found in (A). Abbreviations: Ec, *E. coli* strains; Hi, *Haemophilus infuenzae*; Pm, *Pseudomonas aeruginosa*; Sf, *Shigella flexneri*; St, *Salmonella typhimurium*, Yp, *Yersinia pestis*.

However, they allow to differentiate between *S. flexneri* and *S. typhimurium* strains and collectively these peptides can provide discrimination to exclude all the remaining bacteria as possible matches. Strain specificities of

peptide sequences derived from more conserved proteins, like glyceraldehyde-3-phosphate dehydrogenase, chaperone Hsp70 and the chain elongation factor EF-Tu are increasingly lower. However, each of them provides a unique discriminating power and collectively they provide important information about the similarity between the investigated strain and bacteria represented in the database. Hence, the inspection of the histogram shown in Fig. 4A allows to draw important diagnostic conclusions about the analyzed sample. Namely, the analyzed sample contains peptides derived from a bacterium with a high similarity to *E. coli*, *S. flexneri* and *S. typhimurium*, and moderate similarity to *Y. pestis* and *P. multocida.* The observed bacterial similarities, as indicated in Fig. 4, originate from the sequence homology among these and many other proteins that are coded by homologous genes and have significant matches in related bacterial species. Hence, data shown in Fig. 4 reflect these genomic similarities and consequently are highly redundant.

To address the issue of an accepted peptide possibly matching with proteomes of different bacteria, a simple deconvolution filter can be applied for bacterial identification scoring. The ABOid© algorithm assumes that a bacterium with the highest number of matching peptides is deemed to be the most likely candidate of a true match (English et al., 2003). Giving these circumstances, deconvolution can be performed iteratively by selecting the highest scoring bacterium and filtering out peptides assigned to this microorganism from histogram bins associated with all remaining bacteria, which generates a new peptide-matching histogram. The result of the application of this filtering procedure to *E. coli* data is shown in Fig. 4B. This new histogram clearly suggests that almost all degenerate peptides found in proteomes of bacteria other than *E. c*oli K-12 are smaller subsets of the tryptic peptides found in *E. coli* K-12. A clear identification of *E. coli* K-12 is represented in this new histogram. Genomic similarities, inferred from histograms like the one shown in Fig. 4A, between analyzed bacterium and database microorganisms are quite obvious for microbiologists because all these high scoring bacteria are taxonomically close. However, with the expanded database such histograms are much more complex to interpret, especially in the case of mixtures of bacteria or during analysis of environmental samples. Therefore, to allow for more systematic and objective analysis of MS/MS data the following conceptual approach was devised and implemented.

Amino acid sequences of peptides inferred from MS/MS data can be traced to specific nucleotide sequences located on bacterial chromosomes that represent fragments of genes encoding the expressed proteins. In the approach described above, we did not count assignments to particular proteins or genes but to a collection of genes that are a part of a specific bacterial chromosome. However, the contemporary taxonomy of bacteria, based on molecular methods allow for hierarchical grouping of microorganisms that reveal their phylogenetic relationships on the basis of nucleotide sequence similarities. All bacteria represented in our database belong to 12 phyla (see the legend to Fig. 5) that were demarcated mainly on the basis of sequence similarities of genes encoding their small subunit ribosomal RNA (SS rRNA). Hence, in the first step we assume that chromosomes of all bacteria representing a given phylum form a sort of a 'super chromosome'. Therefore, initially we consider only matches to such super chromosomes. The application of this approach to *E. coli* data presented above gives a histogram of assignments (Fig. 5) that can be easily interpreted and forms a basis for an automated, computerized classification scheme used by our ABOid© software.

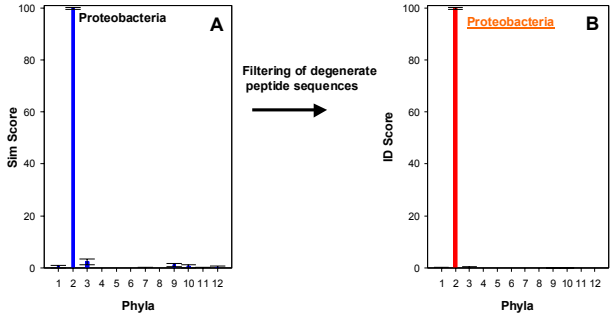

Fig. 5. Percentages of accepted peptides from the analysis of an *E. coli* K-12 protein digest matched to proteomes created by combining database organisms in accordance with their taxonomic position on the phylum level (averages +/- SD; n = 15). [Phyla: 1-*Cyanobacteria*, 2-*Proteobacteria*, 3-*Firmicutes*, 4-*Planctomycetes*, 5-*Thermatogae*, 6-*Aquificae*, 7-*Chlamydiae*, 8-*Deinococcus-Thermus*, 9-*Bacteroidetes/Chlorob*i group, 10-*Actinobacteria*, 11-*Spirochaetes*, 12-*Fusobacteria*] (A) Peptide sequence assignments passing the filter at the 90% probability level; (B) the corresponding identification plot obtained by filtering out the degenerate sequences found in (A).

The relatively high resolving power of this analytical method allows also for the direct analysis of mixtures of microorganisms as documented in Fig. 6. In this case, a bacterial mixture composed of *E. coli* K-12 and *B. subtilis* cells (2:1, w:w) was investigated by 1D HPLC-MS/MS analysis of a protein extract digest. There were 800 unique peptides detected from the SEQUEST analysis of the MS/MS spectra. The graphical representation of assignments to phyla is shown in the upper center of Fig. 6 and clearly documents the presence of *Proteobacteria* and *Firmicutes* in analyzed sample.

In the second step the algorithm separately analyzes assignments matched to both divisions on the class level. Therefore, on this level it considers "super chromosomes" composed of sequences grouped into classes. In our database *Proteobacteria* are represented by 4 classes

(alpha, beta , gamma and delta/epsilon), while *Firmicutes* by three (*Bacilli, Clostridia* and *Mollicutes*). From the histograms of assignment (Fig. 6) it is clear that analyzed mixture is composed of gamma-*Proteobacteria* and *Bacilli*.

In the subsequent steps ABOid© analyzes the assignments on the level of orders, families and genera and, in the case presented in Fig. 6, provides an indication that the analyzed sample was composed of a mixture of organisms that can be placed as the closest relatives of *B. subtilis* and *E.coli* – K12 strains.

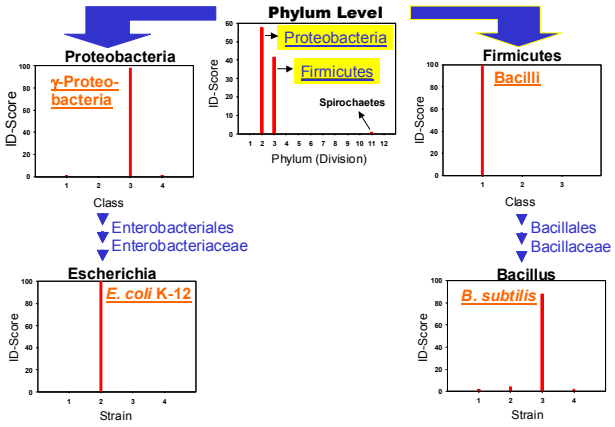

Fig. 6. Schematic representation of the hierarchical classification scheme used for analysis of a bacterial mixture composed of *E. coli* K-12 and *B. subtilis* cells (2:1), w/w) protein extract digest that matched to proteomes of bacteria in the database arranged in accordance to their phylogenetic position (for explanation, see text).

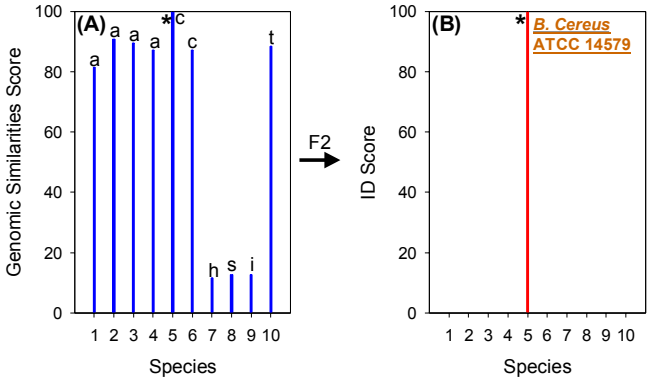

Fig. 7. Percentages of accepted peptides from the analysis of a B. cereus ATCC 14579 protein digest matched to proteomes of database bacteria from the family *Bacillaceae* (A). (B) the corresponding identification plot obtained by filtering out the degenerate sequences found in (A). Abbreviations: a, *B. anthracis* strains (species 1-4); c, *B. cereus* strains (5,6), h, *B. halodurans* (7); s, *B. subtilis* (8); i, *Oceanobacillus* iheyensis (9); t, *B. thuringiensis* (10). For the number assigned to a particular strain – see Fig. 9.

The classification power of our method relies on the diagnostic usage of sequence differences between taxa to identify unknown isolates. However, the survey of named database species for partial sequences identified during the analysis of MS/MS spectra allows for much deeper insight into the relatedness among bacteria by revealing the cluster structure of bacterial strains even on the subspecies level. For example, the 15 min long LC-MS/MS analysis of peptides from a *B. cereus* strain allowed us to identify 136 sequences with an average length of 14 amino acids that can be matched to 10 strains from the family *Bacillaceae* (Fig. 7A). The histogram of assignments indicates that the whole set of them matches a proteome of *B. cereus* ATCC 14579, while only different subsets match the other *Bacillaceae* strains. Hence, the removal of degenerate sequences (Fig. 7B) strongly suggests that the analyzed bacterium is the closest relative of the strain 14579. All these accepted sequences can be traced to 53 different proteins encoded by genes located along the circular chromosome of this bacterium that is shown schematically in Fig. 8 in the linear format. In our procedure any two strains are scored

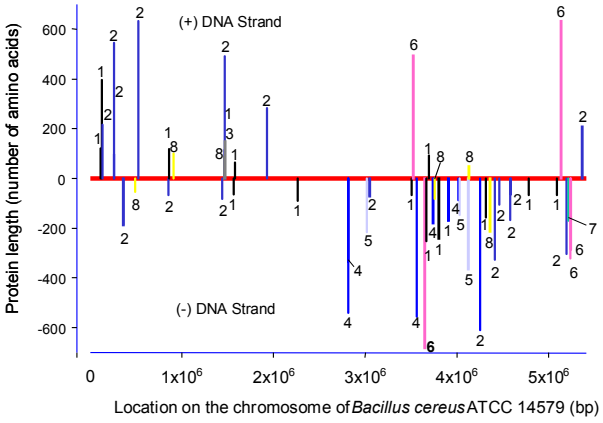

Fig. 8. Locations of genes encoding proteins that contributed peptides identified during LC-MS/MS analysis of a protein extract from B. cereus ATCC 14579 on its circular chromosome. Vertical axis indicates sizes of these proteins expressed as a number of amino acids. Functional categories of proteins: 1-information processing; 2-cellular processing; 3- coenzyme and cofactor metabolism; 4-membrane transport; 5-amino acid metabolism; 6-energy metabolism; 7-nucleotide metabolism; 8-hypothetical.

as different for a sequence segment even if they differ by one non-synonymous substitution on the genomic level. Therefore, the scores for all peptide coding segments can be treated as variables in 136-dimensional space and subjected to cluster analysis. The results of a complete linkage analysis of this data set are presented in Fig. 9 as a horizontal hierarchical tree plot. The Euclidean distances between revealed clusters are expressed as percentages relative to the maximum distance in the plot.

In this case the maximum distance has been found between a cluster formed by *B. halodurans*, *O. iheyensis* and *B. subtilis* and the remaining strains. The latter cluster (marked "z") aggregates strains classified as a *B. cereus sensu lato* (or *B. cereus* group). However, this cluster is composed of two sub clusters, marked as "x" and "y", and both of them include *B. cereus* strains. Nevertheless, the analyzed *B. cereus* strain is assigned as the closest relative of the strain 14579 (marked with an asterisk) that forms a separate cluster ("y") with a *B. thuringiesnsis* strain. Hence, we can conclude that the analyzed strain is different than other *B. cereus* strains, for example 10987, that show closer similarity to *B. anthracis* strains.

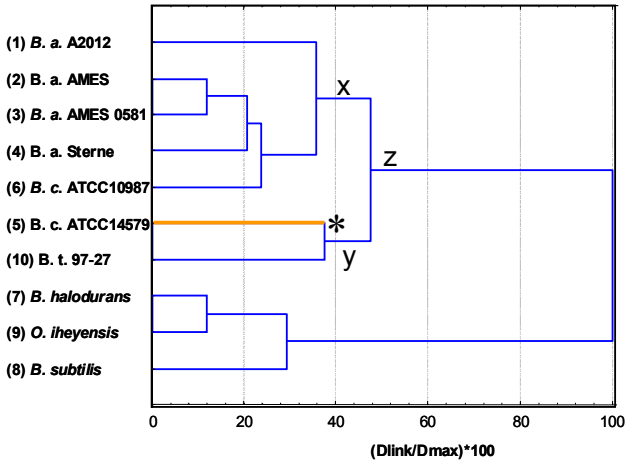

Fig. 9. Graphical visualization of hierarchical clustering of the *Bacillaceae* strains generated by analysis of Euclidean distances in the 136-dimensinal space of identified peptide sequences.

4. CONCLUSION

A new method for fast classification and identification of bacteria based on proteomic analysis of tryptic peptides derived from extracted cellular proteins has been described and the software implementing the identification procedure has been developed. This algorithm utilizes genomic similarities and phylogenetic relationships of database organisms for automated classification and identification of bacteria. The criteria for identifying organisms and assignments to groups were based on a hierarchy of genomic similarities between an analyzed organism and bacteria in the database arranged into groups using phylogenetic relationships. These similarities were calculated using genomic information (equivalent to a few thousand codons) revealed by a set of the most probable amino acid sequences derived from product spectra of peptide ions generated by MS/MS analysis.

REFERENCES

Blattner, F. R.; Plunkett, G., 3rd.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B. and Shao, Y.,1997**:** The Complete Genome Sequence of *Escherichia coli* K-12, *Science*, **277**, 1453-1462.

Aebersold, R.and Goodlett, D. R., 2001: Mass spectrometry in Proteomics, *Chem. Rev.,* **101**, 269 – 296.

Pershing, D. H., Tenover, F. C., Versalovic, J., Tang, Y-W., Unger, E. R., Relman, D. A. and White, T. J., 2003:Molecular Microbiology: Diagnostic Principles and Practice, ASM Press, Washington, D.C.

Chen, W.; Laidig, K. E.; Park, Y.; Park, K.; Yates, J. R., III; Lamont, R. J.and Hackett, M., 2001: Searching the *Porphyromonas gingivalis* Genome with Peptide Fragmentation Mass Spectra, *Analyst*, **126**, 52-57.

Harris, W. A. and Reilly, J. P., 2002: On-Probe Digestion of Bacterial Proteins for MALDI-MS, *Anal Chem.,* **74**, 4410-4416.

Yao, Z-P., Alfonso, C. and Fenselau, C., 2002: Rapid Microorganism Identification with on-Slide Proteolytic Digestion Followed by Matrix-Assisted Laser Desorption/Ionization Tandem Mass Spectrometry and Database Searching, *Rapid Commun. Mass Spectrom.,* **16**, 1953-1956.

Warscheid, B.and Fenselau, C., 2003: Characterization of Bacillus Spore Species and Their Mixtures Using Postsource Decay with a Curved-Field Reflectron, *Anal. Chem.,* **75**, 5618-5627.

Dworzanski, J. P.; Snyder, A. P.; Chen, R.; Zhang, H.; Wishart, D. and Li, L., 2004: Identification of Bacteria Using Tandem Mass Spectrometry Combined with a Proteome Database and Statistical Scoring, *Anal. Chem.,* **76**, 2355-2366.

Saltzberg, S. L.; Delcher, A. L.; Kasif, S. and White, O., 1998: Microbial Gene Identification Using Interpolated Markov Models, *Nucleic Acids Res.,* **26**, 544-548.

Eng, J. K.; McCormack, A. L.and Yates, J. R., 3rd, 1994: An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database, *J. Am. Soc. Mass Spectrom.,* **5**, 976-989.

Keller, A.; Nesvizhskii, A. I.; Kolker, I. and Aebersold, R.,2002: Empirical Statistical Model to Estimate the Aaccuracy of Peptide Identifications Made by MS/MS and Database Search, *Anal. Chem.*, **74**, 5383-5392.

English, R. D.; Warsheid, B.; Fenselau, C.; Cotter, R. J., 2003: *Bacillus* Spore Identification via Proteolytic Peptide Mapping with a Miniaturized MALDI TOF Mass Spectrometer, *Anal. Chem.,* **75**, 6886-6893.